Appendix A

This document outlines the approach to developing a DUI at NCAR, as described in Chapter 1

The Data Usage Index

This study builds off of a previous development of a consumer sentiment model, the "Data Usage Index" (DUI), developed by Peter Ingwersen and Vishwa Chavan (2011; 2012). The DUI combines fourteen quantitative indicators that capture different ways that data, or digital objects are discovered and accessed in an archival setting (See Table 1 for full description of these indicators). Ingweresen and Chavan proposed the creation of these an index that could track the use of species occurrence records from the Global Biodiversity Information Facility (GBIF), in hopes of both identifying valuable digital objects hosted by, and rewarding or incentivizing contributors to that database..

	Code	Indicator		Explanation
1	s (u)	Searched Records		Number of records searched/viewed (by IP address) in unit
2	d(u)	Download Frequency		Number of downloaded records from unit
3	r(u)	Record Numbers		Number of records in (period; dataset(s); geographical and/or species unit)
4	S(u)	Search Events		Number of different searches (by IP address) in unit
5	D(u)	Download Events		Number of different downloads from unit
6	R(u)	Dataset Number		Number of datasets in (period, geographical and/or species unit)
7	s(u)/ S(u)	Search Density		Average number of searched records per search event
8	d(u) / D(u)	Download Density		Average download frequency per download event
9	d(u) / r(u)	Usage Impact		Download frequency per stored record per unit
10	s(u) / r(u)	Interest Impact		Searched records per stored record per unit
11	d(u) / s(u)	Usage Ratio		Ratio of download frequency to searched records in unit
12	D(u) / S(u)	Usage balance		Ratio of download events to search events for unit (in %)
13	U(u) / r(u)	Usage Score		Ratio of unique downloaded records (U) to record number (in %)

| 14 | I(u) / r(u) | Interest Score | | Ratio of unique searched records (I) to record number (in %) |

Table 1. The Indicators for Ingwersen and Chavan’s Data Use Index

Adapting the DUI to the RDA

Although tailored to GBIF and biodiversity data, Ingwersen and Chavan note that the DUI should be adaptable to any research data archive, but that in doing so, ‘…one needs to take into account the fundamental characteristics of datasets and their usage patterns’ within that domain (2011, p. 2). By taking into account "fundamental characteristics" of datasets, Ingwersen and Chavan mean what are commonly thought of as the structural properties of a dataset- its encdings, format, size, etc., and "usage patterns" are then the functional properties of a dataset - how it is accessed, or the way it is made discoverable within an archival. So another way of putting Ingwersen and Chavan's instructions might be this: "The DUI is adaptable to new archival contexts given an understanding of the structural and functional properties for the digital objects being indexed." Thus, our study of these two elements, the structural and functional properties, can serve a dual function: We can better understand the make-up of the

Structural and Functional Properties of ICOADS

In adapting the Data Usage Index from biodiveristy data to climate data, I first looked for structural and functional properties of climate data that were unique to this type of data, and then tried to identify similarities between specific climate datasets so as to create a set of indicators that were generalizable enough for the purpose of a data usage index for all climate data hosted by the RDA.

In the RDA setting, most users do not browse, or search for data in the same way that users of GBIF might look for a set of records based on their geolocation. The need for global data in Climate related work means that instead of location and smaller aggregates of specific evidential records (i.e. fossils as the occurrence of a species in a particular place at a particular time), the coverage, scale of a grid and the reliability of a long-standing dataset is much more important (Parsons, 2010). This means that instead of creating personalized collections of records,

This means that regular, normal and repeated consumption of a single data product is a norm within climate scienece. Because updates to a dynamic dataset (such as ICOADS) are made by a file corresponding to a date (such that one file might contain all weather observations of a field station for August, 2013) I hypothesized that I could capture fluctuations in consumption of a dataset by looking at the amount of data (in files) downloaded by user per download session [^complete].

Datasets in this archive are not typically discovered through generic searching, or through queries made within the RDA's search interface. Instead, many datasets are discovered through reputation, and direct communication within a community of practice (Worley and Jacobson, 2009). Therefore, instead of following Ingweresen and Chavan's use of query logs to understand interest in the dataset, the number of times a user accessed the landing page of a dataset was used as a measure of its popularity and interest.

Thus, the popularity or interest indicator, and the download frequency indicator are the major adaptations of the DUI for climate science data. (See Table 2 below for full description of indicators).

	Code	Indicator	Explanation
1	uu(ds)	Unique Users	Unique users that downloaded data during a time window
2	n(ds)	Number of Datasets	Number of Datasets assigned DS number by RDA
3	f(ds)	Files DS	Number of files in Dataset per time window
4	d(ds)	Download Frequency	Total number of files downloaded per time window
5	hp(ds)	Homepage Hits	Home Page Hits of Data Set per time window
7	d(ds) /uu (ds)	Download Density	Average number of files downloaded per unique user
8	d(ds) / f (ds)	Usage Impact	Total number of downloaded files over total files in dataset
9	d(ds) / hp(ds)	Usage Balance	Files downloaded by number of homepage hits per time window
10	hp(ds) / f(ds)	Interest impact	Total homepage hits per number of files in dataset
11	hp(ds) / uu(ds)	Secondary Interest Impact	Total Homepage hits over unique users
12	ss(ds) / d(ds)	Subset Ratio	Subset requests over total number files downloaded

Table 2: The Indicators for the RDA's Data Use Index

To demonstrate the generalizability of the DUI being proposed two additional earth science datasets were selected to test the DUI's indicators. Overall, this pilot project included the analysis of ICOADS (ds093.0-6); a set of global observational data (ds540), and a popularly analyzed data product from a numerical weather prediction center (ds083). (See Appendix X for full explanation of the two "control group" datasets).

The RDA's access logs for each of these datasets, ranging from January 1, 2010 - June 30, 2012, were then retrieved the relevant indicator data were extracted. Each indicator score was then combined into a one-month time window. (For instance, the total number of ICOADS files downloaded from Aug 1- 31st 2011 were used to create a score titled "Download frequency", such as indicator four in Table 2.) One month time windows were used following the practices of Ingwersen and Chavan (2011) Table 3 shows two, one-month scores for each dataset.

Indicator	ds540.0-1 : 3/2011	ds540.0-1 : 7/2012	ds083.2 : 03/2011	ds083.2: 05/2012	ds093.0-3 : 3/2011	ds093.0-6 : 07/2012
Unique Users	46	45	987	976	88	272
Download Frequency	264	373	374962	335422	3528	23739
Files DS	433	473	22221	25504	195616	277642
Homepage Hits	685	588	6749	6907	1655	3534
Subset Requests	145	35	n/a	42	175	791
Download Density	5.73913043	8.28888888	379.900709	343.67008	40.0909090	87.275735
Usage Impact	0.60969976	0.78858351	16.8736485	13.151740	0.01803528	0.0855021
Interest Impact	1.58198614	1.24312896	0.30371145	0.2708202	0.00846043	0.0127286
Download Ratio	2.59469697	1.57640750	0.01799915	0.0205919	0.46910430	0.1488689
Usage Balance	0.38540146	0.63435374	55.5581567	48.562617	2.13172205	6.7173174
Subset Ratio	0.54924242	0.09383378	n/a	0.0001253	0.04960318	0.0333207
Datasets	2	2	1	1	3	6
Secondary Interest Impact	14.8913043	13.0666666	6.83789260	7.0768442	18.8068181	12.992647

Table 3. Indicator scores for ds540; ds083; and ds093.6

After an initial analysis of these indicator scores, the consumption of each dataset was further investigated by typing users based on the way that a dataset was accessed - either through a command line interface or through a graphic user interface. I hypothesized that typing access in this way would lend insight as to what kind of users a dataset had over time, and how innovations in curation or archival functionality affected their access to these datasets. The two user types are referred to as:

The same one-month time periods were then used to analyze trends between these two types of users (See table 4).

	Code	Indicator	Explanation

1	uu(ds)	Unique Users	Unique users that downloaded data during a time window
1a	uu-p(ds)	Unique Users: Programmatic	Unique users that accessed data programmatically
1b	uu-as(ds)	Unique Users: Assisted	Unique users that accessed data via GUI or RDA Service
4	d(ds)	Download Frequency	Total number of files downloaded per time window
4a	d-p(ds)	Download Frequency: Programmatic	Files downloaded programmatically
4b	d-as(ds)	Download Frequency: Assisted	Files downloaded by Assisted users
5	hp(ds)	Homepage Hits	Home Page Hits of Data Set per time window
5a	hp-dl (ds)	Homepage Hits: Direct	Home Page Hits of Data Set per time window by users with direct link
5b	hp-q (ds)	Homepage Hits: Query	Home Page Hits of Data Set per time window by users link from an indexed list or retrieved by search
8	d(ds) / f (ds)	Usage Impact	Total number of downloaded files over total files in dataset
8a	d-ad(ds) / f(ds)	Usage Impact: Programatic	“
8b	d-as(ds) / f(d)	Usage Impact: Assisted	“

Table 4: Programmatic and Assisted Users of of ds540; ds083; and ds093.6

End notes

- The RDA is a repository of atmospheric and oceanographic observational data, weather prediction model output, gridded analyses and reanalyses, climate model output, and satellite derived data that has been curated by staff in the Computational and Information Systems Laboratory (CISL) at NCAR since the mid 1970's (Jacobs and Worley, 2009). The total volume of the RDA's holdings is approximately 1.3 Petabytes. Most data served by the RDA are dynamic; many datasets are routinely updated, and new datasets or derivative products developed from existing holdings are added each year. The RDA is unique amongst earth science data archives in that it serves almost exactly the same amount of data as it stores- meaning that the RDA last year contained about 1.3 Petabytes of data, and in total it also served users, located all over the world, about 1 Petabyte of data.
- Over the last thirty years staff at the RDA have developed a robust graphical user interface (GUI) allowing for unique subsets of ICOADS data that are not "enhanced" (statistically trimmed).The RDA has also exposed a number of metadata fields that are

meant to help users make the dataset, they provide a suite of Fortran software that allows for observations made in the IMMA (International Maritime Meteorological Archive) format to be read and used in a variety of applications, and they also provide a global archive of year-month observations.

- The distinction between measuring data use, and calculating indicators of "data that are probably used" is subtle but important- the former requiring evidence of how and why people use data in conducting their research, which can only be reliably investigated with expensive, time consuming qualitative methods. The latter suggests that we can assemble usage indicators that are inferential- that is, they likely typify how data are used, or even give some insight as to the popularity or importance of dataset to a given community of users, but they are not direct evidence of use any more than a pageview of an online journal is evidence that the user actually read the article on that page. See (Weber, 2013) ASIS&T Bulletin for a longer discussion.